Semantic Concept Co-Occurrence Patterns for Image Annotation and Retrieval

Linan Feng, *Student Member, IEEE* and Bir Bhanu, *Fellow, IEEE*

Abstract—Describing visual image contents by semantic concepts is an effective and straightforward way to facilitate various high level applications. Inferring semantic concepts from low-level pictorial feature analysis is challenging due to the semantic gap problem, while manually labeling concepts is unwise because of a large number of images in both online and offline collections. In this paper, we present a novel approach to automatically generate intermediate image descriptors by exploiting concept co-occurrence patterns in the pre-labeled training set that renders it possible to depict complex scene images semantically. Our work is motivated by the fact that multiple concepts that frequently co-occur across images form patterns which could provide contextual cues for individual concept inference. We discover the co-occurrence patterns as hierarchical communities by graph modularity maximization in a network with nodes and edges representing concepts and co-occurrence relationships separately. A random walk process working on the inferred concept probabilities with the discovered co-occurrence patterns is applied to acquire the refined concept signature representation. Through experiments in automatic image annotation and semantic image retrieval on several challenging datasets, we demonstrate the effectiveness of the proposed concept co-occurrence patterns as well as the concept signature representation in comparison with state-of-the-art approaches.

Index Terms—Community detection, contextual information, hierarchical co-occurrence patterns, image concept signature

1 Introduction

Representing images by semantic concepts instead of visual features remains a challenging problem. Generating semantic descriptors manually is not feasible due to the ever-growing number of image collections. Current machine intelligence and statistical learning techniques for inferring semantic concepts from low-level features struggle in bridging the semantic gap [1]. However, many image-based applications such as retrieval, annotation, recommendation, indexing and ranking, require an effective semantic representation of images. There is a growing need in automatically inferring concepts from visual properties by learning the correspondence from loosely labeled data.

Semantic concepts cover not only objects that are used in many recognition tasks but also topics at the semantic levels beyond single objects. These higher semantic level could be a scene (e.g., beach), an event (e.g., commencement), and a piece of knowledge (e.g., how to drive a car). A simple form of contextual information is the co-occurrence frequencies of groups of concepts that appear across images with similar scenes. Visual co-occurrence can be quite important in providing semantic cues in inferring concepts compared to other conceptual and perceptual models [42] such as the WordNet distance [40] which is built upon semantic similarity. It has been shown [41] that co-occurrence of concepts could consolidate the appearance of each concept in an image. For example, if "horse" and "windmill" forms a co-occurrence pattern, then the probability of occurrence of "horse" could be reinforced by a strong confidence of "windmill" inference, while the occurrence of "zebra" could be rejected because it has a weak co-occurrence with "windmill". Discovering co-occurrence patterns of semantic concepts is an essential step to encode contextual information into the individual concept inference.

There are two main **contributions** of this paper. *First*, we propose a novel approach to discover the co-occurrence patterns in a network structure where the nodes represent semantic concepts and edges represent co-occurrences. The significance of the co-occurrence relationship between two concepts is denoted by the edge weight. A common property that has been discovered in many networks is the *community structure* property, which is the partition of network nodes into groups (communities) with highly inter-connected nodes (more edges with higher weights), and nodes belonging to different groups being sparsely connected (fewer edges with lower weights). Inspired by the theories in network analysis, we discover the concept co-occurrence patterns by identifying communities in a network. We adopt modularity optimization [2] based approach to uncover *hierarchical community structure* which naturally reflects the co-occurrence patterns at different closeness levels. The idea of hierarchical community structure and co-occurrence patterns is illustrated in Fig. 1. To our knowledge, our work is the first attempt to explore concept co-occurrences from the network analysis point of view. Detailed experimental support for this contribution is provided in Sections 4.4, 4.5.1, and 4.6.1.

Second, we introduce a novel random walk based approach to utilize the discovered co-occurrence patterns to generate "concept signature", a new image representation using high-level semantic concepts to assist in image

- *L. Feng is with the Department of Computer Science and Engineering, University of California, Riverside, CA 92521. E-mail: fengl@cs.ucr.edu.*
- *B. Bhanu is with the Center for Research in Intelligent Systems, University of California, Riverside, CA 92521. E-mail: bhanu@cris.ucr.edu.*

Manuscript received 22 Oct. 2013; revised 9 Dec. 2014; accepted 3 Aug. 2015. Date of publication 16 Aug. 2015; date of current version 11 Mar. 2016. Recommended for acceptance by D. Ramanan.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2015.2469281

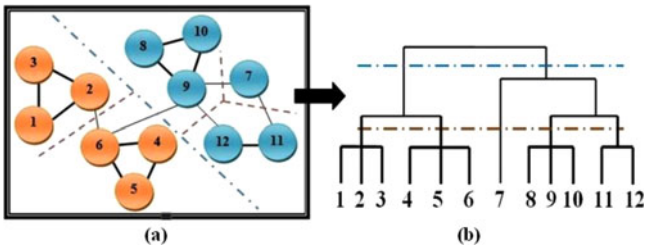


Fig. 1. An illustration of (a) a network of nodes representing the semantic concepts and the edges representing the co-occurrence relations, and (b) the discovered corresponding hierarchical community structure from the network that shows concept co-occurrence patterns at different levels.

annotation and retrieval. The hypothesis here is that the probability scores of uncertain semantic concepts in the concept signature that are generated from the inference model can be promoted or weakened based on the reliably inferred members in a co-occurrence pattern. We demonstrate that our concept signature representation can be very useful in annotation and retrieval of complex scene images. Experimental results in the proposed application scenarios on popular benchmark datasets show clear gains from co-occurrence patterns as compared to other baseline approaches with/without exploiting concept correlations. Detailed experimental support for this contribution is provided in Sections 4.5.2 and 4.6.2.

The remainder of this paper is organized as follows: Section 2 summarizes the related work and contrasts it with the contributions of this paper. Section 3 describes the proposed framework and various algorithms. Section 4 presents experimental results and performance evaluation. Finally, Section 5 gives the conclusions of the paper.

2 Related Work

In the following, we review those approaches that are most relevant to our research along three directions: **(i)** Models that investigate concept correlations as contextual information for image based applications, **(ii)** Image semantic descriptors and **(iii)** Network analysis approaches for detecting communities.

2.1 Semantic Concept Co-Occurrence Models

The approaches based on co-occurrence models for concept inference in complex scene images have gained an increasing popularity [4], [5], [6]. In [40], [41], pairwise concept co-occurrence has been integrated into the concept categorization framework by using a co-occurrence matrix. These approaches have several advantages over standard concept inference techniques, for example, incorporating semantic context compensates the ambiguity of concept visual appearance. However, the matrix of the co-occurrence has an inevitable pairwise constraint on the relationship.

Several recent works explore multi-concept learning/detection techniques for automated image annotation that aim to model the co-occurrence information among concepts/annotations. A simple way is to rank the related concepts based on their co-occurrence relations in the training set and use the ranked relations to refine the annotation results. The idea is similar to collaborative filtering (CF) [50] used by the recommender systems [51]. CF has been introduced in image retrieval [52] to collect the relevance feedback co-occurrences. One of the challenges for CF is the data sparsity problem where the image-concept matrix used for collaborative filtering could be extremely large and sparse in a large image dataset. Matrix-factorization (MF) [53] has been found to be accurate and scalable to address the sparsity problem in CF. By introducing the non-negative constraint into the MF process (NNMF), Zhou et al. [54] proposed a CF method for concept correlation estimation, and Liu et al. [55] presented a framework for semi-supervised multi-label learning using NNMF. Li et al. [56] proposed a multi-correlation probabilistic matrix factorization model to seamlessly estimate the image-concept, image-image and concept-concept correlations simultaneously. Desai et al. [60] examine spatial co-occurrence statistics and incorporate it as contextual relations. Our approach in this paper is significantly different from the above works in discovering the co-occurrence patterns of concepts of any size by detecting the patterns as social communities in a network structure.

To learn more reliable contextual relationships among the semantic concepts, multi-task learning [57] has been introduced for hierarchical image annotation which requires the incorporation of concept ontology. Fan et al. [58] constructed the concept ontology using semantic and visual similarity of concepts, in an attempt to explore the inter-concept correlations and to organize the image concepts in a hierarchy. Multi-task learning is adopted to overcome the problem of intra-concept visual variations. Bourdev et al. [59] presented a hierarchical concept learning framework by incorporating concept ontology and multi-task learning to enhance the image classification performance with a large concept vocabulary. Our approach not only avoids the pairwise constraint, but also, more interestingly, it relies more on the contextual relationships (co-occurrences) rather than the perceptual relationships (concept ontology) that are used in the multi-task learning frameworks [57], [58], [59].

Another problem in existing approaches is that one concept cannot be shared among co-occurrence groups. For example, the method proposed in [16] attempts to discover the co-occurrence between objects by learning a tree structure using Chow-Liu algorithm based on pairwise mutual information. But in their tree structure a concept at the root can only have relationships with the children in its subtree, and cannot have any relationship with the nodes in its siblings' subtrees. Also the same concept cannot be duplicated and shared between subtrees. For instance in their tree structure, "sky" only has a connection with "mountain" but not with "tree" and "road" which may not be true in many cases. In contrast, our proposed approach addresses the overlapping of concepts explicitly.

One of the drawbacks in existing work [40], [42] is the dataset limitation. To find the co-occurrence relationships between objects, these papers do not use strongly labeled data. Instead, they rely on outside sources such as Google Sets, WordNet and Word Association. However, these sources usually do not consider the visual co-occurrence, namely, they are purely based on text or semantic meaning similarity. For example, Google Sets leverage the word

co-occurrence on web pages *without* considering the actual observations in images. WordNet is purely based on the semantic meaning similarity to determine the distance between concepts. It does not reflect the actual co-occurrence property in images. However, in our work, we use the datasets for which the labels are given only when the corresponding concept are observed in an image.

Many algorithms for detecting concept correlations used graph models which is close to our idea. Probabilistic graph models that focus on batch-mode concept detection are proposed in [14]. Correlation of concept co-occurrence and relative spatial locations in images are captured by a tree model in [16]. Besides the positive correlations, they also modeled the negative relationships in the tree structure.

2.2 Image Semantic Descriptors

Many papers in computer vision adopt semantic representations for multimedia understanding and scene analysis, and for applications such as semantic based image annotation and retrieval [43], [44], [45]. Berg et al. [48] automatically generate natural language sentences from *gist* features at different image sizes. Since their final goal is to generate sentence description for an image, the image descriptor is only used in an intermediate procedure and it is still based on visual features which will have a gap between the semantic meaning of images. Unlike these descriptors, our image signature representation focuses on mid-level semantic concepts that are not too general (e.g., *forest*, *desert*) and not too specific (e.g., *palm tree, NIKE shoes*) and addresses the semantic gap problem explicitly.

Ali et al. [47] generate semantic descriptions for images in the form of sentence annotations. Instead of predicting sentence from an image directly, they provide an intermediate step to compute the meaningful triplet (object, action and scene). They name the set of triplets as *meaning space*. The idea of finding the most matched triplet from the meaning space for an image is similar to our concept of finding co-occurrence patterns from the network structure. However, the meaning space is used only as an intermediate step for predicting the sentences, it is not used as a semantic descriptor for comparing image similarity as in our work.

Attribute representation has become a trend in image classification [62], [63] and visual recognition [64], [65] due to its intuitive way in interpreting images and cross-category generalization [66]. Unlike visual words, semantic attributes are sharable discriminative visual properties that are machine-detectable and human nameable (e.g., "square" as a shape property, "silk" as a texture property, "has wing" as a sub-component property, and "can fly" as a functional property). One advantage of semantic attributes is that they naturally bridge the gap between low-level visual features and high-level concepts. In other words, semantic attributes can be used to answer not only "how" two images are similar in a human interpretable way [67], but also "why" an image is identified to belong to a specific category [68]. Attributes are also used frequently in multimedia retrieval as an intermediate semantic description [45]. As compared to the attribute-based representations, our concept signature is generated from the inference models combined with a refinement process that utilizes the co-occurrence information of the concepts.

The most similar image descriptor to our concept signature is the *Object Bank* representation [46]. However, there are several differences. First, the Object Bank representation is computed on grids over an entire image but the grids usually do not fully match the object geometry. Instead, we compute concept signature for each segmented salient region and the signatures are concatenated to form the final image descriptor. Second, each object in the bank is selected based on the occurrence frequency across different datasets. However, we do not consider cross-dataset concept occurrence because an indoor concept may not have frequent occurrences in an outdoor dataset. Third, object bank is used to address the scene classification and object recognition tasks while our concept signature is used for image annotation and semantic image retrieval.

2.3 Network Structure and Community Detection

Network structure has drawn great attention in analyzing relationships between objects. Network structures are proposed in [12], [17] as context graph where individual concepts are nodes and the edges between them are weighted by multi-modal similarities, and random walk algorithms along the context graph are used to refine the detected concepts. In [13] a visual conceptual network (VCNet) is constructed based on the proposed *Flickr distance* to represent the conceptual correlation. A very common property in many realistic complex networks such as social networks and biological networks is known as the *community structure*. Traditional algorithms for detecting community structure can be categorized into graph partition based methods [18], hierarchical clustering algorithms which can be further classified into agglomerative (e.g., [19]) and divisive (e.g., [20]) algorithms, spectral algorithms [21], modularity-based methods [2], and dynamic algorithms [22].

As compared to the above related work in Sections 2.1 to 2.3, the key contributions of this paper are summarized towards the end of 1.

3 Technical Approach

The flowchart of the proposed framework with the contributions (see 1) is shown in Fig. 2. To leverage the contextual information, we only deal with the images with multiple concepts. In order to acquire a reliable individual concept detector, the training images are labeled at the object level, i.e., the concepts are given with the minimum bounding rectangles, and the visual features are extracted regionally. In the semantic sense, a pool of concepts is collected from the training set as the vocabulary to construct the co-occurrence network (described in Section 3.1.1) for concept co-occurrence pattern detection (Section 3.1.2).

The semantic concepts are used to build probabilistic models for inferring the correspondence between a semantic concept and the relevant visual features (Section 3.2). We use both generative and discriminative models, for comparison, as individual concept detectors to discover the semantic concepts in the test images.

Concept signature is introduced as visual and semantic description of images with its elements obtained from the individual concept inference results (described in Section 3.3). With the help of the uncovered concept co-occurrence

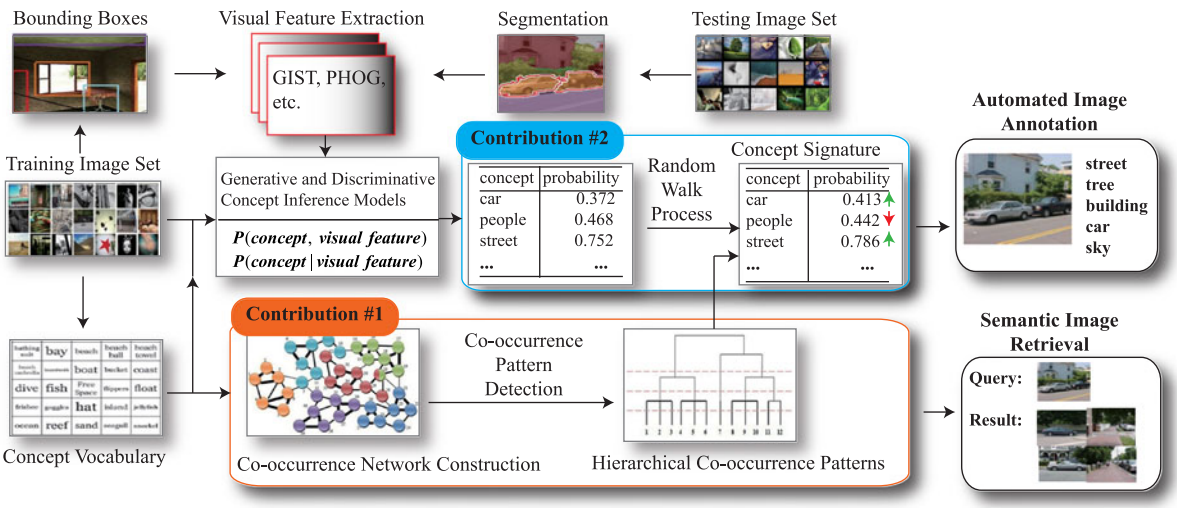


Fig. 2. The flowchart of the proposed concept inference framework. The contributions are: (i) a co-occurrence pattern detection method that effectively explores hierarchical correlations among semantic concepts, (ii) random walk based approach to refine the concept signature representation based on detected concept co-occurrence patterns.

patterns, the concept signature is further refined to approach the ground-truth labels through a random walk process. The effectiveness of the proposed framework is evaluated experimentally in 4 for automatic image annotation and concept-based image retrieval applications. Table 1 summarizes the definition of symbols used in 3.

3.1 Construction of Co-Occurrence Network and Pattern Detection

3.1.1 Co-Occurrence Network Construction

In this section, we discuss the representation of various co-occurrence relationships among different semantic concepts. As the number of concepts is large and the relationship among them tend to be complex, we model them by a network structure. In this paper, we name such a network of concepts as *Concept Co-occurrence Network* (CCN). Let $G = (V, \omega)$ represent a network structure, where each edge $e \in E$ is assigned with a positive weight $\omega(e)$ corresponding to its importance in the network. Let $\Phi = \{c_1, c_2, \ldots, c_m\}$ be the concept vocabulary in the training image set, where m is the total number of unique concepts annotated to the images that the system is attempting to detect. Let $T = \{t_1, t_2, \ldots, t_n\}$ denote the training image set with size n. The CCN is constructed by associating each concept c_i with a node v_i in G. Concepts with textual and visual appearances in the same media resource are likely to have co-occurred and should be linked together by an edge in E.

The edge weight is determined by three types of co-occurrence measure, namely, *global semantic co-occurrence* measures, *global visual co-occurrence* measure and *local visual co-occurrence* measure. First, We evaluate the global semantic co-occurrence by the normalized Google distance [11] (NGD). NGD is proposed to compute the pairwise conceptual distance by counting the number of web pages containing the query concept returned by Google search engine. NGD is intrinsically a co-occurrence measure that explores the co-occurrence of words from on-line textual documents assuming a global meaning of words. Second, for global visual co-occurrence measure, we adopt Flickr based normalized tag distance [12] (NTD) measure. NTD treats the tag list associated with each image in a role similar to the web page in NGD and it calculates the conceptual distance in the same way. Since tag lists indicate visual co-occurrences of concepts in social media resources, it is very intuitive to use NTD to reflect the global frequency of concept co-occurrences. Finally, we apply automatic local analysis [15] (ALA) to identify local visual co-occurrence of concepts in a particular image dataset denoted as the training set in order to capture the local co-occurrence property in the specified image collection. The motivations and the usefulness of the three measures are summarized in Table 2.

Algorithm 1. CCN Construction

Input: Training image set T with n images, a vocabulary Φ with m individual concepts
Output: Constructed concept co-occurrence network $G = (V, \omega)$

1 Initialize a $m \times m$ concept adjacency matrix A for recording edge weights with every element set to 0.;
2 Measure the global semantic co-occurrence between each pair of concepts $\{c_i, c_j\}, i \in 1, \ldots, m, j \neq i$ by normalized Google distance [11]:
$NGD(c_i, c_j) = \frac{max\{logG(c_i), logG(c_j)\} - logG(c_i, c_j)}{log\Omega - min\{logG(c_i), logG(c_j)\}}$;
3 Measure the global visual co-occurrence by normalized Tag distance [12]: $NTD(c_i, c_j) = exp\frac{max\{logF(c_i), logF(c_j)\} - logF(c_i, c_j)}{log\Psi - min\{logF(c_i), logF(c_j)\}}$;
4 Measure the local visual co-occurrence by automatic local analysis [15]: $ALA(c_i, c_j) = exp(-\Delta)$, where
$\Delta = \frac{\sum_{t_k \in T} x_{ik} \times x_{jk}}{\sum_{t_k \in T} x_{ik} \times x_{ik} + \sum_{t_k \in T} x_{jk} \times x_{jk} - \sum_{t_k \in T} x_{ik} \times x_{jk}}$;
5 Combine the three measures into the final co-occurrence significance and assign the value to element $A(c_i, c_j)$;
6 $A(c_i, c_j) = \eta_1 \cdot NGD(c_i, c_j) + \eta_2 \cdot NTD(c_i, c_j) + \eta_3 \cdot ALA(c_i, c_j)$. In our setting we put equal importance on the three measurements, so $\eta_i = \frac{1}{3}$;
7 Traverse all the elements in A, add c_i as node, connect two nodes c_i, c_j with edge weight according to the value of A_{ij};

TABLE 1
Definition of Symbols Used in Section 3.1

Symbols	Definitions
$G(V, \omega)$	The constructed concept co-occurrence network with V and E representing the node and edge sets separately, and ω denoting the edge weight.
v_i	The ith element in the node set V.
Φ	The vocabulary of semantic concepts in this work.
m	The number of semantic concepts in vocabulary Φ.
c_i	The ith element in the concept vocabulary Φ.
T, t_i	The training image set and the ith element.
n	The number of images in the training set.
$A_{m\times m}$	The adjacency matrix used to record the edge weights in G, $A(c_i, c_j)$ denotes the weight of the edge connecting concepts c_i and c_j.
$H_{m\times n}$	The association matrix, $h_{ik} = 1$ if concept c_i appears in image t_k in the training set and 0 otherwise.
G(c)	The number of pages containing concept c reported by Google search engine.
G(c_1, c_2)	The number of pages containing concepts c_1 and c_2.
Ω	The number of pages indexed by Google.
F(c)	The number of images containing concept c in Flickr.
F(c_1, c_2)	The number of images containing both concepts c_1 and c_2 in Flickr.
Ψ	The number of images indexed by Flickr.
$x_{i,k}$	Equals 1 if concept c_i appears in training image t_k, 0 otherwise.
η_1, η_2, η_3	The weights set to evaluate the importance of each co-occurrence measure, $\sum_{i=1}^{3} \eta_i = 1$.
C	The community detected in the network structure.
Q_C	The modularity measure of community C.
ΔQ	The modularity gain acquired when the community structure changes.
k_i	The summation of edge weights attached to node v_i in the network.
$k_{i,C}$	The summation of edge weights where the edges are connecting node i to the nodes in community C.
Γ	The half of the summation of all the edge weights.
δ	The delta function used in computing the modularity.
Σ_{in}	The summation of edge weights inside community C.
Σ_{out}	The summation of edge weights that link to the nodes outside community C.
Λ_c	The visual variation of semantic concept c.
R_c, $\lvert R_c\rvert$, r_c^i	Training region set containing concept c, the size of the set and the ith element.
$R_{\bar{c}}$, $\lvert R_{\bar{c}}\rvert$, $r_{\bar{c}}^j$	Negative training region set of c, the size of the negative set and the jth element.
f_{R_c}	The mean of the feature vectors of the regions in R_c.
$f_{r_c^i}$	The feature vector of ith region r_c^i in R_c.
$f_{r_{\bar{c}}^j}$	The feature vector of jth region $r_{\bar{c}}^j$ in $R_{\bar{c}}$.
Z	The dimension of the above feature vectors.
D_{χ^2}	The Chi-square distance between two feature vectors.
$\mathcal{G}$	The function that generates the prototype vector.
g	The prototype vector generated from a region.
w	The weight vector in the SVM objective function.
b	The bias vector in the SVM objective function.
e_1, e_2	The constants for controlling the relative influence of the two competing terms in the SVM function.
h	The hinge loss function in the SVM objective function.
CS	The concept signature descriptor.
s_{c_i}	The confidence score of concept c_i in the signature.

The motivation for using the three co-occurrence measures is that they can complement one other. NGD uses the entire World-Wide-Web as the dataset which is known to be the largest on earth. The contextual information is given by billions of independent persons of knowledge, thus, it can overcome the limitation in the scope of the concepts represented in image datasets. However, NGD does not involve any visual information in the distance calculation, and co-occurred concepts in the textual documents may have zero probability to appear in the real life images (e.g., concepts from *science-fiction novels*). Therefore, visual co-occurrences are analyzed to decrease the ambiguities arisen from texts. Global visual co-occurrences from community-contributed web image collections, e.g., Flickr, are represented by the rich tags as metadata. However, it cannot accommodate the changes to the training dataset. i.e., images and concepts that are added or removed from the original dataset. Local visual co-occurrence can contribute to dynamic dataset, thus, it is reasonable to be considered. The steps for constructing the CCN are described in Algorithm 1.

3.1.2 Co-Occurence Pattern Detection

Finding the co-occurrence patterns of the interconnected nodes corresponds to uncovering community structures from the randomness of the network topology which is close to graph clustering or partitioning problem. However the problem is computationally intractable. Recently *modularity* has been used as a criterion for determining the effectiveness of the detected communities, and at the same time

TABLE 2
The Summarization of the Usage and Motivation for Adopted Co-Occurrence Measures

Co-occurrence Measure	Usage & Motivation
Normalized Google Distance (NGD)	Captures the global semantic co-occurrences. The number of semantic concept co-occurrences in a local dataset is far below than what is generated by the massive web users. For example, there are 434 million concepts/annotations found from web images [3]. NGD can actually reflect the confidence that two semantic concepts can co-occur among online textual resources.
Normalized Tag Distance (NTD)	Captures the global visual co-occurrences. NGD assumes concept relationships only depend on semantic co-occurrences in the text field which cannot guarantee the existence of these co-occurrences from the visual perspective (i.e., the presence in the images). NTD treats the tags that are associated with the images as the general semantic concepts used in NGD and calculates the co-occurrence in the same way as NGD. It strengthens the visual co-occurrences between concepts.
Automatic Local Analysis (ALA)	Captures the local visual co-occurrences. NGD and NTD utilize the global information that is out of the scope of a local dataset. However, global co-occurrences may not exactly match the local co-occurrence in an image collection. Therefore, ALA is introduced to strengthen the local visual co-occurrences.

it can serve as an objective function to maximize. In this paper we adopt *modularity optimization* paradigm to address the problem and propose a method based on Newman-Girvan modularity [2] optimization. The modularity measures the quality of a partition by comparing the link density of nodes inside a community with the links to the outside nodes. Usually high values of modularity suggests good partitions. In the case of weighted network, we define the modularity of community C as:

$$Q_C = \frac{1}{2\Gamma}\sum_{i,j}\left[A_{ij} - \frac{k_i k_j}{2\Gamma}\right]\delta(ID_i, ID_j) \tag{1}$$

Typically modularity score is in the range of $[-1, 1]$, and in practice a value greater than 0.3 indicates a significant community. The modularity is calculated over all the pairs of nodes in the network, where ID_i and ID_j are their community IDs, $\delta(ID_i, ID_j) = 1$ if $ID_i = ID_j$ for two nodes v_i and v_j, otherwise $= 0$. We consider iteratively merging the nodes into a hierarchical community structure with different levels of resolution by maximizing the modularity gain at each iteration. The *modularity gain* of moving an outside node v_i into a community C is evaluated by:

$$\begin{aligned}\Delta Q = &\left[\frac{\Sigma_{in} + k_{i,C}}{2\Gamma} - \left(\frac{\Sigma_{out} + k_i}{2\Gamma}\right)^2\right] \\ &- \left[\frac{\Sigma_{in}}{2\Gamma} - \left(\frac{\Sigma_{out}}{2\Gamma}\right)^2 - \left(\frac{k_i}{2\Gamma}\right)^2\right]\end{aligned} \tag{2}$$

Please see Table 1 for the definitions of symbols. Algorithm 2 is given for detecting the hierarchical concept co-occurrence patterns (communities) in a network. The runtime of the algorithm for co-occurrence pattern detection is $O(|V|(|E| + |V|))$ where $|E|$ is the number of edges and $|V|$ is the number of nodes in the network. The algorithm iteratively generates a hierarchical community structure with different resolutions, in other words, the communities of individual concepts, and the communities of communities. To point out, our algorithm addresses the share of nodes problem between communities explicitly.

Algorithm 2. Concept Co-occurrence Pattern Detection

Input: Co-occurrence network built from Algorithm 1
Output: Hierarchical concept co-occurrence patterns

while *Positive Modularity Gain can be achieved* **do**
  Partitioning phase;
  foreach c_i *in the vocabulary* ($i = 1, \ldots, N$) **do**
    Assign Node n_i represents c_i in the network;
    Label n_i with community tag C_i;
  Each node will have an unique community tag after above step;
  while *Positive Modularity Gain can be achieved* **do**
    foreach n_i *in the network*
      Remove n_i from its original community C_i;
      foreach *neighboring community* C_j of n_i **do**
        Add n_i to C_j;
        Calculate modularity gain ΔQ (eq.(2)) after changing the community structure;
        if $\Delta Q > 0$ **then**
          Let C_{old} and C_{new} denote the original community and new community of node n_i;
          Compute modularity scores $Q_{C_{old}}$ and $Q_{C_{new}}$ by eq.(1);
          if $Q_{C_{old}} >= 0.3$ *and* $Q_{C_{new}} >= 0.3$ **then**
            n_i is shared by both communities;
            Split n_i into n_i and $n_i^{'}$;
            Add n_i into C_{old} and add $n_i^{'}$ into C_{new};
            Copy the edges of n_i that are incident to other nodes for $n_i^{'}$;
          else if $Q_{C_{old}} < 0.3$ *and* $Q_{C_{new}} >= 0.3$ **then**
            Change the community tag of n_i from C_{old} to C_{new};
          else
            n_i stays in the original community;
      else
        n_i stays in the original community;
  Coarsening phase: generates the hierarchical structure;
  Replace the nodes in the same community detected from the above steps as a single node;
  Replace the edges between the nodes in two adjacent communities by a single edge with summed edge weights;
  Represent edges in the same community as a self-looped edge with weight equal to the sum of the internal edge weights;

3.2 Concept Occurrence Inference Models

We integrate the detected concept co-occurrence patterns for individual concept inference. We use probabilistic inference models to build the correspondence between concepts and regional visual features from training data. The outputs of the model running on testing images are vectors of concepts with corresponding probabilities scores of the occurrence. We name this vector representation as *concept signature* which captures both the semantic and visual information about images.

Individual concept inference is the baseline and key factor to the overall performance although we demonstrate later that it can be improved by utilizing the co-occurrence patterns. In order to compare the effect of the baselines, we implement two individual concept inference models based on generative and discriminative training.

The *generative model* is built by jointly estimating the probability of visual and semantic representations. Suppose T is the training set of annotated images and R is the set of corresponding segmented regions, and let r be an element of R. We specify the process of feature vector generation and vector quantization as an integrated function $\mathcal{G}$, with $g = \mathcal{G}(r) \in \mathbb{R}^J$. Each image t in T can be represented as a set of regions $r_t = \{r_1, r_2, \ldots, r_n\}$ along with the corresponding concept from the set $\{c_1, c_2, \ldots, c_n\}$. Given an image region r, *first*, we model the probability of obtaining concept c by sampling from a multinomial distribution $P_{\mathcal{M}}(c|r)$ that will split probability mass among multiple concepts. Subscript $\mathcal{M}$ represents the multinomial distribution [61]. *Second*, we model the relation between a region r in the training set and a possible prototype vector g as a distribution $P_{\mathcal{R}}(r|g)$. *Finally*, when given a region r from the unknown set, we model the probability of getting a prototype vector g by sampling from a distribution $P_{\mathcal{G}}(g|r)$. For an unknown region r_i from a test image, the probability of observing c_j is given by the joint probability:

$$P(r_i, c_j) = \sum_{r_t \in R_{c_j}} \left\{ P(r_t) \cdot P_{\mathcal{M}}(c_j|r_t) \cdot \left\{ \sum_{g_t} P_{\mathcal{R}}(r_i|g_t) \cdot P_{\mathcal{G}}(g_t|r_t) \right\} \right\} \quad (3)$$

We assume that the training set is sufficient to cover all possible instances of the region-concept pair in the test set. The larger the size of the training set, the more correct knowledge about the generative model that we can obtain. The details of the probabilities in eq. (5) are given in [61].

The *discriminative model* is created by an ensemble of instance-SVMs for each concept where the idea is similar to [49]. For each concept, the positive instances are the regions containing that concept and the rest are negatives. We first train a separate linear SVM classifier for each positive instance of a given concept with the negatives. For each positive instance with feature $f_{r_c^i}$ of concept c, and the negative set $R_{\bar{c}}$ with instance feature $f_{r_{\bar{c}}^j}$, the weight vector w are learned by optimizing the convex objective:

$$\upsilon(w, f_{r_c^i}, b) = ||w||^2 + e_1 h(w^T f_{r_c^i} + b) + e_2 \sum_j h\big(-w^T f_{r_{\bar{c}}^j} - b\big) \quad (4)$$

where h represents the hinge loss function $h(x) = (0, 1 - x)$ which permits hard-negative mining to find the small subset of dominating negative support vectors from $R_{\bar{c}}$. For a test region, the instance-SVM classifiers of a concept are first applied. The outputs from individual classifiers are fused by weighted averaging to generate the final concept score. The weight w attached to each single classifier is determined by adaptive linear neural network (ALNN) in a validation process. We give the details of the implementation in [61].

3.3 Concept Signature and Its Refinement

We propose concept signature as image descriptor. Concept signature is a vector in which each entry contains a tuple of concept and its occurring probability from the inference model. Compared to other image descriptors, concept signature: 1) records both the visual and semantic information of an image, thus, image can be compared and retrieved based on high-level semantic concept similarity, which we denote as *concept-based image retrieval* in this paper. 2) has a very simple form, therefore, it can lower the memory cost for storing large image collections and decrease the computational costs. 3) can keep all the concept occurrence probabilities which can be revised later on when individual concept inference accuracy is improved.

We refine the original scores in the concept signature in a *re-ranking* manner formulated as a random walk process over the contextual co-occurrence patterns. Suppose the hierarchy has L levels, we set the lowest level that contains the semantic concepts as level-1 and the highest level as level-L. Assume initially concept c_i has occurring score s_{c_i} given by the inference model, and let $lowest(c_i, c_j)$ denote the function to compute the level of the lowest superordinate (common ancestor) between c_i and c_j. In the kth updating iteration, the score s_{c_i} is refined by the random walk process:

$$s_{c_i}^k = \alpha \sum_{c_j \neq c_i} s_{c_j}^{k-1} \cdot \frac{lowest(c_i, c_j)}{L} + (1 - \alpha) \cdot s_{c_i}^{k-1} \quad (5)$$

We set α to 0.5 which means the effects from its own score and the scores from neighboring concepts are treated equally. The scores are updated recursively until all the scores converge. Eq. (5) can strengthen the scores of concepts in more closely related patterns and weaken the more isolated ones. Finally, we give Algorithm 3 for generating image concept signature and random-walk refinement.

Algorithm 3. Concept signature refinement

Input: Testing image set
Output: Refined concept signature representation for each testing image

1 **foreach** *Image T in the testing set* **do**
2 Detect the salient regions $r_1, \ldots, r_m$ by mean shift based segmentation [25];
3 **foreach** *Salient region r_i* **do**
4 Apply the inference models defined in eq.(3) or eq.(4);
5 Compute the original regional signature $CS_{r_i} = ((c_1, s_{c_1}), \ldots, (c_n, s_{c_n}))$;
6 Compute the intermediate image-level signature by $CS_I = \frac{1}{m} \sum_{i=1}^{m} CSr_i$;
7 Obtain the final image concept signature by random walk based refinement (eq.(5));

4 Experimental Results

4.1 Image Datasets and System Parameters

4.1.1 Image Datasets

- The **LabelMe** [26] dataset is a collection of 72,852 images containing more than 10,000 concepts. We use a subset which contains 10,000 images and 2,500 concepts. The raw images have different resolutions (e.g. $2{,}560 \times 1{,}920$, $1{,}600 \times 1{,}200$, 256×256, etc.). In this paper, we use the resolution of $1{,}600 \times 1{,}200$ downloaded from the website by using the Toolbox provided by the dataset creators.
- The **Scene Understanding (SUN'09)** [16] dataset contains 12,000 images and more than 5,800 concepts covering a variety of indoor/outdoor scenes. The total number of annotated labels is 85,456 which results in an average of seven labels per image. The images are collected from multiple sources (Google, Flickr, Altavista, LabelMe) and are labeled by a single annotator using the LabelMe tool. The labels are manually verified for consistency.
- The **Outdoor Scene Recognition (OSR)** [27] dataset has 2,682 images with 520 concepts across eight outdoor scene categories: coast, forest, highway, inside-city, mountain, open-country, street, tall-building. All the concepts are labeled with corresponding bounding boxes manually.

The selected datasets have the following advantages compared to other datasets (e.g., TinyImages [28], MSRC [29], Caltech-101 [30]): **(i)** All the datasets present complex scenes containing multiple concepts in a single image which is suitable for exploring the concept co-occurrence correlations. **(ii)** Compared to the general and specific terms defined in the *synonym set* in WordNet (e.g., "mammal", "tool", "geological formation") and used by ImageNet (e.g., "coconut tree", "ocean floor", "Davy Jones"), most of the concepts are at the intermediate level of semantics (e.g., "tree", "sea", "people") which are more relevant to Folksonomy-style tags used in daily life. **(iii)** The datasets have a large number of concepts that cover a great majority of object categories. **(iv)** The bounding boxes for the concepts are available in standard XML format which can be easily parsed by programs (e.g., the open source tool TinyXML [39] used in our framework).

4.1.2 System Parameters

The weighting parameter ω (Section 3.1.1) is set to $1/3$ for the three measures. The modularity threshold Q_C (Section 3.1.2) is set to 0.3, and the weight parameter α in the random walk process (eq. (7)) in this paper is set to 0.5. All the parameters are set empirically and they are kept constant for all the experiments reported in this paper.

4.2 Visual Features

We extract visual features locally from the regions enclosing the concepts defined by minimum bounding rectangle (MBR). For test images, the features are extracted from the MBRs of the segmented salient regions. The features are:

- **Color GIST** feature [27] is computed on 4×4 grids over the concept bounding box. The MBRs are resized to 32×32 (we do not maintain the aspect ratio) and then the orientation histograms are calculated at three scales with 8, 8 and 4 bins.
- The **pyramid of histogram of oriented gradients (PHOG)** feature [31] is computed by following steps: 1) extract the Canny edges in the concept bounding box, 2) quantize the gradient orientation on the Canny edges (from 0 to 180 degree) into 20 bins, 3) Four spatial pyramid levels are used ($1 \times 1, 2 \times 2, 4 \times 4, 8 \times 8$). Each level is used in an independent kernel.
- **PHOG with oriented edges** [32] considers the direction (0 to 360 degree divided into 40 bins) of the salient Canny edges. We use four-level spatial pyramid.
- The **pyramid of Shechtman and Irani self similarity** feature [33] is computed at every 5 pixels and quantized into 300 clusters using k-means, and then the histograms are calculated at three levels.
- The **bag of visual words** feature [32] is obtained by first computing the SIFT descriptors [34] at the interest points detected by Hessian-Affine detector [35], and then quantizing them into a vocabulary of visual words with the size of 1,000. Finally, a sparse histogram is generated based on the visual words.

4.3 Applications and Evaluation Criteria

4.3.1 Application 1: Automatic Image Annotation

The goal is to predict concept occurrences for an image from a given concept vocabulary. The predictions are then used to annotate the image based on the rank of the probability scores. Most existing approaches for AIA neglect the co-occurrence patterns among concepts and annotate the concepts individually. In our framework, the concepts ranked as top-M in the refined concept signature based on the inferred probability scores are used as the annotations. An alternative way with unfixed annotation length is to use all the annotations with scores passing certain threshold.

4.3.2 Application 2: Concept-Based Image Retrieval

For a given query, we compute the similarity to the database images based on the concept signature representation using the Earth Mover's Distance (EMD) [36] as the distance metric. Given two concept signatures $\mathbf{p}$ and $\mathbf{q}$, the EMD is defined as: $\mathrm{EMD}(\mathbf{p}, \mathbf{q}) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} o_{ij} d(p_i, q_j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} o_{ij}}$, where o_{ij} denotes the flow and it follows the constraints of the scores in the concept signature and $d(p_i, q_j)$ is the pre-defined ground distance between each pair of individual concepts. In our setting, we use the reciprocal of the edge weight in the co-occurrence network as the measure of ground distance. EMD measures the least amount of work to completely transfer one signature into another, it is calculated by linear programming [36].

4.3.3 Evaluation Criteria

- Automatic image annotation: The performance is evaluated by *Top-M* $F_{0.5}$ measure, *Top-M* F_1 measure and *Precision* measure for a given annotation length M. In our case, we set M to 5. F_β measure is defined

TABLE 3
Pairwise Co-Occurrence Scores for Example Concept Pairs by Using NGD, NTD, ALA and the Combination of the Three

Pairwise Co-occurrence Scores (normalized to [0,1])				
Concept Pairs	NGD	NTD	ALA	Combined
mountain-tree	0.551	0.597	0.448	0.532
sky-cloud	0.713	0.825	0.629	0.722
road-car	0.533	0.614	0.687	0.611
street-building	0.429	0.475	0.512	0.472
sand-sea	0.217	0.483	0.359	0.353
ground-grass	0.261	0.385	0.297	0.314
person-terrance	0.097	0.152	0.219	0.156
door-window	0.483	0.509	0.411	0.468
rock-hill	0.202	0.317	0.384	0.301
sun-land	0.215	0.158	0.278	0.217
river-boat	0.343	0.416	0.357	0.372
sidewalk-sign	0.294	0.187	0.371	0.284
field-fence	0.482	0.359	0.411	0.417
wall-staircase	0.128	0.119	0.274	0.174
curb-streetlight	0.213	0.319	0.307	0.280

as $(1+\beta^2)\cdot(P\cdot R/\beta^2 P+R)$, where P is the averaged per-image precision and R is the averaged per-image recall. When we set β to 0.5, we put more emphasis on precision than recall. The reason is that the ground-truth annotation length is usually more than the fixed length we used for most of the images. Therefore, even we get all the annotations correct, we still cannot reach the best recall score. Instead, we look for better performance by considering the true positives in the total five annotations. However, to give more information on the performance, we also provide the results evaluated by standard F_1 measure and $Precision$ measure.

- Image retrieval: The performance is evaluated by the ranks of the relevant images in the returned results. We have five human assessors launched queries using each database image and provide relevance information on the retrieved images. The degree of relevance of a retrieved image is calculated by the total number of assessors who submit "relevant" decision divided by five. Further statistical evaluation relies on the standard image retrieval measure: *Mean average precision (MAP) of top D retrieved images* over all the images. Let D be the retrieved image set and R be the relevant ones with size $|R|$. Given a query Q, the average precision is defined as $AP(Q)=\frac{1}{|R|}\sum_{i=1}^{|R|}\frac{i}{Rank(R_i)}$, and the mean average precision is the averaged AP over all the images.

4.4 Co-Occurrence Pattern Detection Results

4.4.1 Experiment I: Co-Occurrence Measure Study

We apply our co-occurrence pattern detection approach on a network built from the training set of each dataset. LabelMe contains 2,500 individual concepts, SUN'09 contains 5,800 concepts, and OSR has 520 concepts.

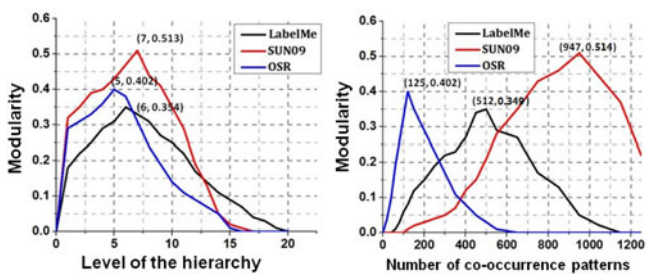
We demonstrate that our combined co-occurrence measure of NGD, NTD, and ALA is more effective than each of the individual measures in co-occurrence network construction as well as co-occurrence pattern detection in the following experiments. First, we compare example pairwise concept co-occurrence scores computed by different measures in Table 3. The scores are averaged over the three datasets and normalized to the range $[0,1]$. Generally, we find the results from NGD, NTD, ALA are more coherent on the pairs with degrees of co-occurrences that are more consistent to human perception (e.g., "mountain-tree", "sky-cloud", and "road-car") than the less consistent ones (e.g., 'sand-sea'", "person-terrance" and "rock-hill"). However, our combined measure is able to reach the maximum consensus among the three. For example, our combined measure is able to leverage the information from NGD and NTD to increase the co-occurrence score of ALA from 0.448 to 0.532 for the pair of "mountain-tree", and is able to use local information from ALA to improve the co-occurrence measure of NGD and NTD for the pair of "wall-staircase". From Table 4 we can observe the effectiveness of using the combined measure in co-occurrence pattern detection evaluated by the modularity score (eq. (1)). Our combined measure gives the best performance in modularity measure from 5th level to 10th level in the hierarchy. The reason for this is that the combined measure can leverage both the global and local co-occurrences as well as utilize both the semantic and visual information.

TABLE 4
Averaged Modularity Scores (Q) from 5th to 10th Level

Modularity Scores				
Datasets	NGD	NTD	ALA	Combined
OSR	0.218	0.259	0.224	**0.275**
SUN09	0.152	0.170	0.143	**0.212**
LabelMe	0.173	0.164	0.139	**0.197**

4.4.2 Experiment II: Impact from the Hierarchy

Fig. 3a shows the change in modularity for different levels of hierarchy in the three datasets. We observe that the maximum of modularity for LabelMe occurs at level 6 with $Q\approx 0.354$, the maximum for SUN'09 occurs at level 7 with $Q\approx 0.513$ and the maximum for OSR occurs at level 5 with $Q\approx 0.402$. This indicates that the individual concepts in SUN'09 have significant community property than OSR and LabelMe, and even appear at lower level of SUN'09 (from level 7 to level 12), the community property is comparatively large compared to the LabelMe and OSR datasets. Fig. 3b shows the correspondence between the number of co-occurrence patterns and the modularity values at



(a) (b)

Fig. 3. (a) Modularity versus level of the hierarchy. (b) Modularity versus the number of co-occurrence patterns.

TABLE 5
Precisions at Different Annotation Lengths by Using Different Co-Occurrence Measures

LabelMe				
Co-occurrence Measure	Pre@1	Pre@3	Pre@5	Pre@10
NGD	0.3393	0.2584	0.2230	0.1276
NTD	0.3752	0.2772	0.2481	0.2025
ALA	0.3806	0.2857	0.2564	0.1847
Combined	**0.4628**	**0.4533**	**0.4279**	**0.3104**
SUN09				
Co-occurrence Measure	Pre@1	Pre@3	Pre@5	Pre@10
NGD	0.3528	0.2693	0.2432	0.1384
NTD	0.3423	0.2537	0.2593	0.1457
ALA	0.3516	0.2714	0.2581	0.1543
Combined	**0.4332**	**0.4233**	**0.4017**	**0.3042**
OSR				
Co-occurrence Measure	Pre@1	Pre@3	Pre@5	Pre@10
NGD	0.3393	0.2584	0.2230	0.1276
NTD	0.3752	0.2772	0.2481	0.2025
ALA	0.3806	0.2857	0.2564	0.1847
Combined	**0.4423**	**0.4323**	**0.4264**	**0.3504**

different levels of the hierarchical community structures. From Fig. 3b we can compute the average number of concepts in the co-occurrence patterns by dividing the total number of concepts by the number of co-occurrence patterns. LabelMe has approximately five concepts averaged over all the co-occurrence patterns at the maximum modularity point, similarly, SUN'09 has six concepts and OSR has four concepts. Note the averaged number of concepts in the co-occurrence patterns are consistent with the averaged number of concepts contained in the training images.

4.5 Automatic Image Annotation Results

4.5.1 Experiment I: Co-Occurrence Measure Study

Table 5 presents the precisions obtained for the three datasets at different annotation length ($Pre@1$, $Pre@3$, $Pre@5$, $Pre@10$) by using four co-occurrence measures: NGD, NTD, ALA, and our *Combined*. $Pre@N$ denotes the precision of annotations in the first N words using 60 percent of the dataset for training. Overall, our combined co-occurrence measure achieves the best performance especially when the annotation length is larger than 1. The reason is that for more annotations more co-occurrence information can be utilized. Generally, when the length of the annotation becomes larger, it deteriorates the annotation precision, however, using combined co-occurrence information our proposed measure still can achieve relatively stable performance regardless of the dataset complexity differences. Furthermore, the number of true positives exceeds 30 percent for our co-occurrence measure at the length of ten annotations which implies that at least three annotations on average are correctly given by our approach. Note that, in general, the contribution from local visual co-occurrence, which is adopted by ALA, surpasses the contributions from global semantic co-occurrence and global visual co-occurrence which are adopted by NGD and NTD, respectively. This demonstrates that each dataset has unique co-occurrence patterns which are different from the global ones. However, by introducing the global information, we can actually consolidate the common patterns which may lack enough samples in a local dataset and weaken the unusual patterns.

4.5.2 Experiment II: Annotation Performance

To demonstrate the effectiveness of our proposed framework for the image annotation application, we evaluate the following approaches as shown in Table 6:

- *Baseline-Gen model*: Our generative implementation for individual concept inference unified with concept signature representation served as the base model. (The base model does not include co-occurrence pattern detection and random walk boosting).
- *Baseline-Dis model*: The discriminative version of the baseline-gen model. The other setup is the same as in baseline-gen.
- *conditional random field (CRF)*: The conditional random field based image annotation approach by Xiang et al. [14] that uses the original pairwise co-occurrences from a network structure without hierarchical co-occurrence pattern detection. We re-implemented it to compare it with our hierarchical pattern scheme.
- *Context*: The object detection and localization approach by Choi et al. [16] that is used for image annotation. They introduced a tree-structured context model which is comparable to our network structure and hierarchical patterns. We re-implemented it to compare its performance with our approach.

TABLE 6
The Top-5 $F_{0.5}$ Score and the Standard Deviation (Show in the Parentheses) of Automated Annotation with Different Training Set Sizes

	LabelMe Dataset [26] (%)			SUN09 Dataset [16] (%)			OSR Dataset [27] (%)		
Methods / % of training data	40%	60%	80%	40%	60%	80%	40%	60%	80%
Baseline-Gen	21.85 (3.253)	29.57 (2.857)	32.81 (2.772)	23.44 (3.157)	32.52 (2.684)	35.14 (2.435)	25.81 (2.217)	35.33 (1.936)	40.47 (1.854)
Baseline-Dis	21.51 (2.857)	31.27 (2.864)	33.03 (2.513)	21.73 (2.679)	33.03 (2.324)	35.94 (2.185)	23.33 (1.906)	34.52 (1.873)	39.72 (1.535)
CRF [14]	25.59 (2.095)	33.81 (2.137)	36.04 (2.241)	26.93 (1.958)	35.71 (1.742)	40.15 (1.699)	27.93 (1.732)	38.91 (1.589)	43.93 (1.489)
Context [16]	26.33 (1.964)	34.13 (1.842)	36.23 (1.765)	27.71 (1.753)	36.58 (1.689)	40.81 (1.626)	28.14 (1.541)	38.63 (1.439)	44.18 (1.387)
HCP-Gen (This paper)	28.74 (1.154)	39.92 (1.112)	41.37 (1.037)	29.52 (1.096)	39.11 (0.965)	44.32 (0.854)	28.71 (0.896)	42.64 (0.859)	**48.36** (0.791)
HCP-Dis (This paper)	28.13 (1.032)	40.85 (1.006)	**43.46** (0.987)	28.23 (1.043)	40.71 (0.958)	**45.68** (0.875)	28.47 (0.955)	41.92 (0.890)	47.71 (0.873)
maximum % gain over CRF	12.30%	**20.82**%	**20.59**%	9.61%	**14.00**%	**13.77**%	2.79%	**9.59**%	**10.08**%
maximum % gain over Context	9.15%	**19.69**%	**19.96**%	6.53%	**11.29**%	**11.93**%	2.03%	**10.38**%	**9.46**%

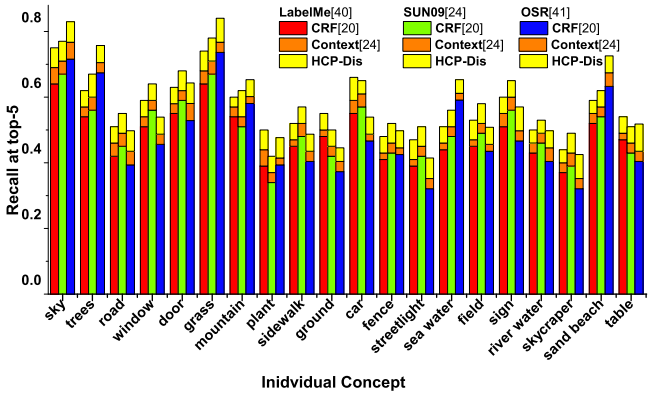


Fig. 4. The recall rate of common concepts in the three datasets.

- *HCP-Gen*: This is our proposed framework integrating generative concept inference, co-occurrence pattern and random walk boosting. HCP refers to hierarchical co-occurrence pattern.
- *HCP-Dis*: A framework with a discriminative concept inference model and everything else is the same as in HCP-Gen.

We evaluate the impact of the training set size by Top-5 $F_{0.5}$ measure averaged over all the testing images. We split the datasets into training and testing sets with three size configurations. For each split configuration we repeated the experiment 10 times by using each of the approaches. Table 6 summarizes the data splits, mean performance and standard deviations. For results on Top-5 F_1 and Precision measures, please see [61]. The tables show that the impact of training set size is obvious and consistent across different datasets. The larger the training set, the better performance can be achieved for all the approaches. Our approach shows clear improvements over the other models reflected by the maximum percent gain (achieved by using HCP-Gen or HCP-Dis). Also, there is a significant performance gain when the training data size exceeds the testing data size for all the three datasets (see the last two columns for each dataset in Table 6. In general, all the approaches require at least 50 percent of the dataset used for training to have reasonable annotation performance. Even the performance of our framework is deteriorated when the training data is under 40 percent.

Next, to analyze the scalability of our approach, we compare the results on the three datasets with increased complexity (OSR $<$ SUN09 $<$ LabelMe) evaluated by the total number of concepts in the datasets and the number of concepts per image. Table 6 shows that generally when the images are complex the performance of the approaches drop. This is demonstrated by the Top-5 $F_{0.5}$ measure (as well as Top-5 F_1 and precision measures [61]). In particular, we observe that our approach achieves better maximum performance gain when the images have higher complexities. For example, LabelMe usually has more than 10 concepts in an image, the maximum performance gain reaches 20.59 percent when the training set contains 80 percent of the images. SUN09 contains on average 5-10 concepts per image, the maximum performance gain is between 11.29 and 14.00 percent. OSR has the least number of concepts in an image, and the maximum gain is the lowest as well

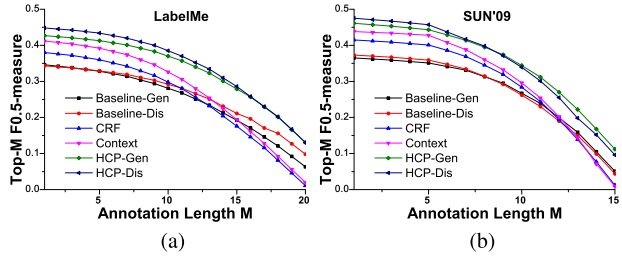


Fig. 5. (a),(b) show the image annotation performance of the approaches applied to the three datasets measured by Top-5 $F_{0.5}$-measure with annotation length $M = 5$.

which is approximately 10.00 percent only. This indicates that our approach is well suited for understanding images with complex scenes. Table 6 also shows that the performance increase by our approach is less compared to other approaches when the images are relatively simple as in the OSR dataset.

We further compare the recall rates at top-5 annotation length obtained by CRF [14], Context [16] and our HCP-Dis approach on selected common concepts across the three datasets. The results are given in Fig. 4. We observe that the contextual information from the three datasets have different effects on individual concept inference. For example, the recall rates for most of the concepts in LabelMe are relatively lower than SUN09 and OSR. The reason for this is that there are more noisy annotations, such as the misspelling and meaningless words in LabelMe from the folksonomy-style annotations, and these noisy annotations deteriorate the co-occurrence pattern detection performance and have adverse impact on the individual concept refinement. OSR dataset has larger recall rates on outdoor concepts while has smaller recall rates on other concepts. We stack the recall rates obtained by different approaches into a single column and we observe that Context [16] (with hierarchy) performs better than CRF [14] (without hierarchy) while our approach always has the highest performance gain on the recall rate. This demonstrates the effect of using hierarchical co-occurrence patterns vs. no hierarchy. Additionally, the recall rates of CRF [14], Context [16] highly depend on the visual consistency of the semantic concepts. For concepts have large intra-concept visual variations (e.g., "road", "ground", "streetlight", and "skyscraper" in Fig. 4), the performance drops greatly especially for CRF which only considers the original pairwise concept co-occurrences. On the other hand, our approach can maintain relatively stable performance which demonstrates the effectiveness of utilizing contextual information obtained from the detected co-occurrence patterns.

Figs. 5a, 5b show the performance comparison based on the Top-M $F_{0.5}$-measure for the three datasets as a function of the annotation length M. As the number of annotations increases, we observe that the performance of baseline approaches, CRF [14] and Context [16] drops faster than our proposed HCP approaches, which demonstrates that our co-occurrence pattern and refinement has a boosting effect on individual concept inference. Further, our approach is more effective in using contextual information than CRF [14] and Context [16] because we explore the correlations of concepts beyond pairwise relationships. We also

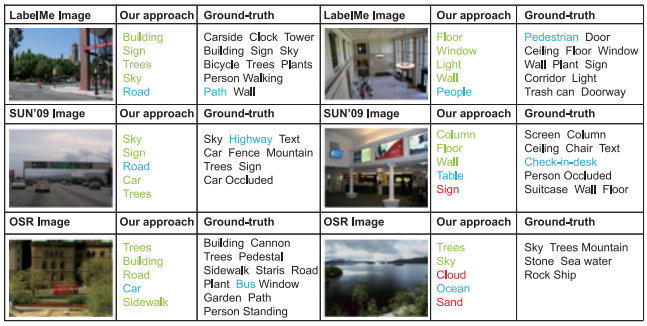


Fig. 6. The annotations for the test images from the three datasets by our approach. They are compared with the ground-truth. Green labels are correctly predicted, red ones are wrongly predicted and blue ones have very close semantic meaning to the ground-truth.

observe that our discriminative model and generative provide approximately the same boost in performance compared to the other approaches. However, HCP-Dis performs better than HCP-Gen for the datasets such as LabelMe and SUN'09 that have more complex scenes and more semantic concepts in a single image. Therefore, we conclude that HCP-Dis has a stronger discriminative power when the number of semantic concepts that share increasingly high visual similarity in an image. Also HCP-Gen can better tolerate the intra-concept visual variation in simple scenes.

Fig. 6 shows the top-5 annotation results for some example images that are produced by our approach. The annotations in green color are the correctly predicted labels and red ones are mistakenly predicted. It is interesting to look at the annotations in blue. These concepts are inferred from the detected individual concepts and co-occurrence patterns. Although they are not exactly the same as the annotations in the ground-truth, but they are close in the meaning for a specific scenario, e.g., "road" and "path" in an "outdoor - street view" scenario, "people" and "pedestrian" in an "indoor - hall" scenario. This shows that our proposed approach can effectively enrich the annotations by considering the scene concepts implicitly contained in the co-occurrence patterns. The refinement capacity of our approach can be seen from the annotation results of the right image in the second row and left image in the last row where the ground-truth concepts "check-in-desk" and "bus" are occluded in the image and the similar concept "table" and "car" are enriched by our proposed refinement strategy. More results are given in [61].

4.6 Concept-Based Image Retrieval Results

4.6.1 Experiment I: Co-Occurrence Measure Study

Table 7 gives the mean average precisions for the datasets at four different sizes of retrieved images (MAP@5, MAP@10, MAP@15, MAP@20) by using four co-occurrence measures: NGD, NTD, ALA, and combined. MAP@N represents the mean average precision of retrieved images in the size of N using 60 percent of the dataset for training. The results in Table 7 show that our combined co-occurrence measure achieves the best performance at all sizes of the retrieved images.

From Table 7 we can observe that the combined co-occurrence measure achieves the best performance and the performance is stable when the size of the retrieved images is less than 15. Even when the size is 20, the combined co-occurrence measure can still have reasonable results in all three datasets. Note that, in general, the contributions from the three individual measures are relatively the same for all sizes of retrieved images. But the boost in MAP values is clear when combining the three measures. This demonstrates that the co-occurrence information from the three measures will compensate each other and it is helpful in learning more accurate concept relationships. Note that the MAP measure is affected by two factors: the difficulty of the dataset and the number of retrieved images. The combined measure can achieve a better MAP compared to the individual measures for all datasets of varying difficulty levels and retrieved image sizes.

TABLE 7
Mean Average Precision for Different Sizes of Retrieved Images by Using Different Co-Occurrence Measures

LabelMe				
Co-occurrence Measure	MAP@5	MAP@10	MAP@15	MAP@20
NGD	0.2564	0.2317	0.1869	0.1003
NTD	0.2616	0.2484	0.2195	0.1574
ALA	0.2543	0.2336	0.1752	0.1249
Combined	**0.2825**	**0.2617**	**0.2797**	**0.1809**
SUN09				
Co-occurrence Measure	MAP@5	MAP@10	MAP@15	MAP@20
NGD	0.2646	0.2334	0.1954	0.1172
NTD	0.2476	0.2318	0.2094	0.1290
ALA	0.2584	0.2027	0.1853	0.1274
Combined	**0.2923**	**0.2898**	**0.2517**	**0.1972**
OSR				
Co-occurrence Measure	MAP@5	MAP@10	MAP@15	MAP@20
NGD	0.2738	0.2418	0.1989	0.1373
NTD	0.2864	0.2529	0.2046	0.1508
ALA	0.2953	0.2591	0.2153	0.1643
Combined	**0.3394**	**0.3004**	**0.2846**	**0.2038**

4.6.2 Experiment II: Image Retrieval Performance

The goal is to show the effectiveness of our concept inference framework for image retrieval task. We implement and evaluate the following approaches for comparison as summarized in Table 3. We also vary the training set size to show its impact on the retrieval performance.

- *Baseline-I*: The content-based image retrieval framework that compares the image similarity by directly computing the euclidean distance between the visual feature vectors as described in Section 4.2.
- *Baseline-II*: The proposed framework integrated with SVM-based individual concept inference. The concept signatures are used directly without refinement by co-occurrence patterns.
- *Semi-supervised graphical model (SSG)*: The approach in [43] uses a latent-tree to find the relationship between semantic concepts. The pairwise relevance is obtained from the graphical model directly. No hierarchical co-occurrence patterns are detected.
- *Hierarchical semantic indexing (HSI)*: The retrieval framework proposed in [38] uses the information

TABLE 8
Mean Average Precision of Top-10 Retrieved Images with Different Training Set Size

Methods / % of training	LabelMe Dataset [26] (%)			SUN09 Dataset [16] (%)			OSR Dataset [27] (%)		
	40%	60%	80%	40%	60%	80%	40%	60%	80%
Baseline-I	7.64 (2.857)	11.93 (3.383)	14.82 (2.754)	9.46 (2.953)	13.53 (2.714)	15.74 (2.906)	15.82 (2.186)	25.37 (2.346)	28.27 (1.974)
Baseline-II	14.43 (2.952)	21.58 (2.742)	26.03 (2.563)	12.71 (3.126)	18.14 (3.064)	22.17 (2.547)	18.93 (2.836)	25.12 (2.914)	29.79 (2.464)
SSG [37]	17.78 (2.532)	24.61 (2.734)	26.94 (2.513)	19.53 (2.631)	25.71 (2.345)	28.45 (2.194)	21.22 (3.126)	28.54 (2.432)	31.82 (1.987)
HSI [38]	17.93 (2.964)	25.17 (2.347)	27.27 (2.146)	19.71 (3.156)	26.15 (2.343)	28.61 (3.134)	21.78 (2.432)	28.78 (1.524)	32.15 (1.768)
HCP-IR (This paper)	18.96 (1.532)	28.97 (1.123)	**32.47** (1.233)	21.06 (1.518)	30.17 (1.425)	**34.22** (1.236)	23.11 (1.435)	33.46 (1.346)	**37.99** (0.983)
% gain over SSG	6.64%	**17.72**%	**20.53**%	7.83%	**19.86**%	**20.28**%	8.91%	**17.23**%	**19.39**%
% gain over HSI	5.74%	**15.10**%	**19.07**%	6.85%	**15.37**%	**19.60**%	6.11%	**16.26**%	**18.16**%

Fig. 7. An example of the top-10 retrieved images by our proposed approach. The retrieved images are ranked based on their semantic distance to the query. The top row shows the correctly retrieved images with street view and the stop sign. In the middle row, the top retrieved images correctly match the bedroom scene represented in the query. And in the last row, the images with a beach scene and people are placed at the top positions.

from generated hierarchical semantic relationships between concepts for comparing image similarity. However, as compared to our work, they do not consider the co-occurrence between concepts, and their concept distance is defined on WordNet.

- *HCP-IR*: Our proposed approach integrated with hierarchical co-occurrence pattern detection and concept signature refinement. We implemented the discriminative model here.

We repeat the split of each of the dataset for 10 times. From Table 8 we can observe that the larger the training set size for all the three datasets, the larger MAP can be achieved by all the approaches. The standard deviations are also given in this table. Baseline-I achieves the worst performance which concludes that traditional content-based image retrieval paradigm is not suitable for retrieving images containing many semantic concepts with a large visual variations. SSG is only marginally better than our Baseline-II approach, for the reason that it only considers the pairwise relationship between individual concepts and the approach is not intended to use images from complex scenes. HSI outperforms SSG while our HCP-IR significantly outperforms both SSG and HSI by 5.74-20.53 percent. This result validates our assumption that the proposed hierarchical concept co-occurrence patterns can boost the individual concept inference. In particular, we can observe that when using only 40 percent of the dataset for training, our method can still achieve comparatively good performance than SSG and HSI. An example of the retrieval results by using our HCP-IR approach with 80 percent training data for the three datasets is shown in Fig. 7. We can observe that the returned images are more semantically related to the scene concept reflected in the query images rather than just visually related. The overall performance of all the approaches decrease when the dataset becomes more complex. However, our approach can maintain a stable maximum gain over SSG [43] and HSI [38].

Figs. 8a, 8b summarize the results for MAP at top-D retrieval results. Our model (HCP-IR) consistently outperforms the other approaches with varying number of retrieved images on the three datasets. This shows the effects of semantic concept correlations and the concept signature descriptor in the context of image retrieval. The results demonstrate that all the components of our framework are essential: (1) detecting individual semantic concepts is important for retrieving images of complex scenes (LabelMe,

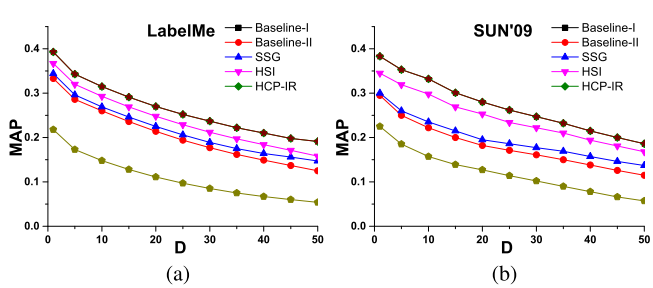


Fig. 8. (a), (b) show the image retrieval performance of the approaches applied to the three datasets measured by Top-D MAP with varied number of retrieved images D.

SUN'09) as Baseline-II is more effective than Baseline-I (directly using low-level features without semantic learning), but for simple scenes (OSR) the two approaches have similar performance (see [61]). (2) learning more sophisticated concept correlation models (HSI, HCP-IR) improves performance over simple pairwise relationships (SSG). We also note a higher precision for OSR than for the other two datasets. This is due to a relatively small number of individual concepts present in the dataset, and therefore, the detected co-occurrence patterns are more significant in more compact forms. For additional results, please see [61].

5 Conclusions

This paper has made a novel contribution to the literature on context-based co-occurrences in computer vision where co-occurrences of concepts are used as contextual cues for improved concept inference. It introduced a framework for individual concept inference and refinement by exploring the concept co-occurrence patterns in images with network community detection algorithms. The framework is evaluated for automated image annotation and concept-based image retrieval tasks using the new concept signature representation. The approach is tested on recent practical datasets and compared with the state-of-the-art methods. The experimental results convincingly show the following: (a) The importance of the hierarchy of co-occurrence patterns and its representation as a network structure, (b) The effectiveness of the approach for building individual concept inference models and the utilization of co-occurrence patterns for refinement of concept signature as a way to encode both visual and semantic information. In the future we will explore the message-passing approach for concept signature refinement and compare it with the random walk based approach.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0905671 and 1552454.

References

[1] H. Ma, J. Zhu, M. R.-T. Lyu, and I. King, "Bridging the semantic gap between image contents and tags," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 462–473, Aug. 2010.
[2] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Phys. Rev. E*, vol. 69, 2004.
[3] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien, "SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation," in *Proc. 20th Int. Conf. World Wide Web*, 2003, pp. 178–186.
[4] S. Hwang and K. Grauman, "Reading between the lines: Object localization using implicit cues from image tags," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 1145–1158.
[5] A. Torralba, "Contextual priming for object detection," *Int J. Comput. Vis.*, vol. 53, no. 2, pp. 169–191, 2003.
[6] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 1271–1278.
[7] Y. Jin, L. Khan, L. Wang, and M. Awad, "Image annotations by combining multiple evidence & WordNet," in *Proc. ACM Multimedia*, 2005, pp. 706–715.
[8] Y. Jin, L. Wang, and L. Khan, "Improving image annotations using WordNet," in *Proc. 11th Int. Conf. Adv. Multimedia Inf. Syst.*, 2005, pp. 115–130.
[9] Y. A. Aslandogan, C. Their, C. T. Yu, and N. Rishe, "Using semantic contents and WordNet in image retrieval," in *Proc. 20th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1997, pp. 286–295.
[10] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.
[11] R. Cilibrasi and P. M. B. Vitanyi, "The Google similarity distance," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 370–383, Mar. 2007.
[12] D. Liu, X. Hua, L. Yang, M. Wang, and H. Zhang, "Tag ranking," in *Proc. 18th Int. Conf. World Wide Web*, 2009, pp. 351–360.
[13] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li, "Flickr distance: A relationship measure for visual concepts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 5, pp. 863–875, May 2012.
[14] Y. Xiang, X. Zhou, Z. Liu, T.-S. Chua, and C.-W Ngo, "Semantic context modeling with maximal margin conditional random fields for automatic image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3368–3375.
[15] R. Baeza-Yates, and B. Ribeiro-Neto, *Modern Information Retrieval*, New York, NY, USA: ACM Press, 1999, pp. 123–129.
[16] M. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting hierarchical context on a large database of object categories," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 129–136.
[17] W. H. Hsu, L. S. Kennedy, and S-F. Chang, "Video search reranking through random walk over document-level context graph," in *Proc. IC Multimedia*, 2007, pp. 971–980.
[18] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, no. 3-5, pp. 75–174, Feb. 2010.
[19] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.
[20] M. E. J. Newman, and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, 2004.
[21] M. Mitrovic and B. Tadic, "Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities," *Phys. Rev. E*, vol. 80, no. 2, 2009.
[22] Y. Hu, M. Li, P. Zhang, Y. Fan, and Z. Di, "Community detection by signaling on complex networks," *Phys. Rev. E*, vol. 78, no. 1, 2008.
[23] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003.
[24] G. Fu, F. Y. Shih, and H. Wang, "A kernel-based parametric method for conditional density estimation," *Pattern Recognit.*, vol. 44, no. 2, pp. 284–294, Feb. 2011.
[25] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.
[26] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, 2008.
[27] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[28] A. Torralba, R. Fergus, and W. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008.
[29] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. 9th Eur. Conf. Comput. Vis.*, 2006, pp. 1–15.
[30] F-F. Li, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 594–611, Apr. 2006.
[31] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. 6th ACM Int. Conf. Image Video Retrieval*, 2007, pp. 401–408.
[32] L. Torresani, M. Summer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proc. 11th Eur. Conf. Comput. Vis.*, 2010, pp. 776–789.
[33] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.
[34] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[35] K.Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.
[36] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[37] B. Siddiquie, R. S. Feris, and L. S. Davis, "Image ranking and retrieval based on multi-attribute queries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 801–808.
[38] J. Deng, A. C. Berg, and F-F. Li, "Hierarchical semantic indexing for large scale image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 785–792.
[39] [Online]. Available: http://sourceforge.net/projects/tinyxml/
[40] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proc. 11th IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.
[41] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.
[42] A. Torralba, K. P. Murphy, and W. T. Freeman, "Contextual models for object detection using boosted random fields," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 1401–1408.
[43] F. X. Yu, R. Ji, M. H. Tsai, G. Ye, and S.-F. Chang, "Weak attributes for large-scale image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 2949–2956.
[44] B. Siddiquie, R. S. Feris, and L. S. Davis, "Image ranking and retrieval based on multi-attribute queries," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 801–808.
[45] M. Douze, A. Ramisa, and C. Schmid, "Combining attributes and Fisher vectors for efficient image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 745–752.
[46] L.-J. Li, S. Hao, E. P. Xing, and F.-F. Li, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.
[47] F. Ali, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 15–29.
[48] O. Vicente, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.
[49] T. Malisiewicz, A. Gupta, and A. A. Efros, "Ensemble of exemplar-SVMs for object detection and beyond," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 89–96.
[50] D. Goldberg, D. Nichols, B. M. Oki, and D. B. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM*, vol. 35, no. 12, pp. 61–70, 1992.
[51] F. Cacheda, V. Carneiro, D. Fernndez, and V. Formoso, "Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems," *ACM Trans. Web*, vol. 5, no. 1, p. 2, 2011.
[52] S. Uchihashi and T. Kanade, "Content-free image retrieval by combinations of keywords and user feedbacks," in *Proc. Image Video Retrieval*, 2005, vol. 3568, pp. 650–659.
[53] X. Luo, M. Zhou, Y. Xia, and Q. Zhu, "An efficient non-negative matrix-factorization-based approach to collaborative-filtering for recommender systems," *IEEE Trans. Ind. Informat.*, vol. 10, no. 2, pp. 1273–1284, May 2014.
[54] N. Zhou, W. K. Cheung, G. Qiu, and X. Xue, "A hybrid probabilistic model for unified collaborative and content-based image tagging," *IEEE Trans. Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1281–1294, Jul. 2011.
[55] Y. Liu, R. Jin, and L. Yang, "Semi-supervised multi-label learning by constrained non-negative matrix factorization," in *Proc. 21st Nat. Conf. Artif. Intell.*, 2006, pp. 421–426.
[56] Z. Li, J. Liu, X. Zhu, T. Liu, and H. Lu, "Image annotation using multi-correlation probabilistic matrix factorization," in *Proc. Int. Conf. Multimedia*, 2010, pp. 1187–1190.
[57] J. Fan, Y. Gao, and H. Luo, "Hierarchical classification for automatic image annotation," in *Proc. Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2007, pp. 111–118.
[58] J. Fan, Y. Gao, and H. Luo, "Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation," *IEEE Trans. Image Process.*, vol. 17, no. 3, pp. 407–426, Mar. 2008.
[59] J. Fan, Y. Gao, H. Luo, and R. Jain, "Mining multilevel image semantics via hierarchical classification," *IEEE Trans. Multimedia*, vol. 10, no. 2, pp. 167–187, Feb. 2008.
[60] C. Desai, D. Ramanan, and C. C. Fowlkes, "Discriminative models for multi-class object layout," *Int. J. Comput. Vis.*, vol. 95, pp. 1–12, 2011.
[61] Supplemental material to accompany this paper by Linan Feng and Bir Bhanu.
[62] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 951–958.
[63] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1543–1550.
[64] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 1778–1785.
[65] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 365–372.
[66] A. Farhadi, I. Endres, and D. Hoiem, "Attribute-centric recognition for cross-category generalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 2352–2359.
[67] Y. Wang and G. Mori, "A discriminative latent model of object classes and attributes," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 155–168.
[68] D. Parikh and K. Grauman, "Relative attributes," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 801–808.

Linan Feng (S'06) received the BSc degree in electrical engineering in 2006 and the ME degree in software engineering in 2009 both from Shanghai Jiao Tong University, China. Since 2009, he has been working toward the PhD degree in computer science at the University of California, Riverside. His research interests are in computer vision, pattern recognition and machine learning, with emphasis on automated image annotation and concept-based image retrieval. He is a student member of the IEEE.

Bir Bhanu (F'95) received the SM and EE degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, the PhD degree in electrical engineering from the Image Processing Institute, University of Southern California, and the MBA degree from the University of California, Irvine. He is the distinguished professor of electrical and computer engineering and cooperative professor of computer science and engineering, mechanical engineering and bioengineering, director of the Center for Research in Intelligent Systems (CRIS), and the Visualization and Intelligent Systems Laboratory (VISLab), University of California, Riverside (UCR). He also serves as the director of US National Science Foundation (NSF) IGERT program on Video Bioinformatics and Interim Chair of the Department of Bioengineering. Prior to that, he was a senior Honeywell fellow at Honeywell Inc. He was the first founding faculty in the Bourns College of Engineering and served as the chair of electrical engineering at UCR. His research interests are Computer Vision, Pattern Recognition and Data Mining, Machine Learning, Artificial Intelligence, Image Processing, Image and Video Database, Biological, Medical, Military and Intelligence applications. He has been the principal investigator of programs from NSF, DARPA, NASA, AFOSR, ONR, ARO and other agencies and industries. He is a fellow of the IEEE, AAAS, IAPR, AIMBE, and SPIE.